Universidad Zaragoza
1542

Trabajo fin de grado

Técnicas de machine learning para la optimización de los resultados del experimento ANAIS-112

Grado en física

Autor:
Jaime Apilluelo Allué

Directores:
Iván Coarasa Casas
Dra. María Lucía Martínez Pérez

Facultad de Ciencias
Departamento de Física Teórica

Junio de 2021

Índice

1. Objetivos

La materia oscura constituye uno de los enigmas más interesantes de la física fundamental actual. A pesar de representar el 85 % de la cantidad de materia total del universo [1], su naturaleza es todavía desconocida. Uno de los candidatos a resolver esta cuestión son los WIMPs (Weakly Interacting Massive Particles), unas partículas no incluidas en el modelo estándar cuya existencia no ha sido todavía demostrada [2]. Estas se caracterizan, entre otras cosas por poseer una masa del orden de 1 GeV - 1 TeV y carga neutra. Además, su interacción con la materia ordinaria estaría en la escala de interacciones débiles, lo que las hace altamente complicadas de detectar. ANAIS-112 es un experimento diseñado para la detección directa de dichas partículas [3]. Para ello, resulta crucial discernir entre una señal correspondiente a la detección de una partícula, y una señal producida por ruido u otro tipo de suceso no originado por la interacción de una partícula en el detector. A tal efecto, se hace uso de diferentes métodos de selección basados en las características de estas señales.

Este trabajo va a centrarse en intentar mejorar los algoritmos de selección de sucesos establecidos en el experimento ANAIS-112 en la región de baja energía. La idea es utilizar métodos de selección de eventos basados en técnicas de aprendizaje automático. Dichos métodos serán puestos a prueba con un conjunto de sucesos escogidos aleatoriamente de las medidas realizadas por ANAIS-112 durante su primer año de funcionamiento, denominado en este trabajo como Población de Test (PT). De esta forma se asegura que la selección de eventos no ha sido premeditada, haciéndola concordar con unos mejores resultados. Se va a trabajar con el paquete TMVA (*Toolkit for Multivariate Data Analysis* [4]) de ROOT [5]. Este incorpora un entorno que facilita el manejo de estos algoritmos centrados en la clasificación multivariante. Concretamente, el algoritmo seleccionado se basa en una extensión de los árboles de decisión clásicos. El programa base, tanto el entrenamiento como la aplicación a los datos, ha sido desarrollado por el doctorando Iván Coarasa, colaborador en el experimento ANAIS-112 [6]. A partir de él se van a analizar diferentes modificaciones, buscando su optimización para conseguir un espectro de fondo a baja energía con la menor cantidad de sucesos de ruido posible, manteniendo en todo caso una eficiencia de detección aceptable.

2. Introducción: el experimento ANAIS-112

El experimento ANAIS-112 se desarrolla en el Laboratorio Subterráneo de Canfranc (LSC), situado bajo el monte Tobazo que supone una cobertura de más de 800 metros de roca. El elemento principal es un detector compuesto por nueve módulos cilíndricos de NaI(Tl) de 12,5 kg (haciendo un total de 112,5 kg). Cada módulo está acoplado a dos fotomultiplicadores (PMT), uno en cada extremo del cilindro. Esta disposición permite establecer un disparo (*trigger*) en coincidencia entre ambos PMT. Este implica que sólo se recojan datos de los sucesos detectados por ambos PMTs dentro de una misma ventana temporal (de 200 ns para ANAIS-112 [3]). Así se realiza un primer criterio de selección de sucesos. Cuando una partícula ionizante atraviesa el material centelleador de NaI(Tl) produce una excitación de sus electrones que, posteriormente pueden recombinarse produciendo una emisión de luz de centelleo que es recogida por los PMT, arrancando, por efecto fotoeléctrico, electrones del cátodo (denominados fotoelectrones, phe) que se amplificarán en el

cuerpo del PMT generando una señal eléctrica, que se leerá como un pulso de voltaje. Este pulso no es instantáneo pues las desexcitaciones del material centelleador tienen una cierta vida media (230 ns en el NaI(Tl) [7]). Así, el pulso correspondiente al centelleo presenta una duración (media) dependiente del material. Además, el área del pulso está relacionada con la cantidad de luz generada, lo que depende tanto del material como de la cantidad de energía depositada por la partícula incidente. En ANAIS-112, esta conversión de energía depositada a fotoelectrones es elevada, de unos 15 phe/keV.

El objetivo del experimento es medir la modulación anual de materia oscura esperada en el ritmo de interacción de esta con el detector debida al movimiento de traslación terrestre, de forma similar a como se midió en el experimento italiano DAMA/LIBRA[8], corroborando o refutando su resultado [9]. En los modelos más aceptados, los WIMPs interaccionan con los núcleos del detector produciendo retrocesos nucleares de energías inferiores a 50 keV. Teniendo en cuenta que los detectores se calibran en energía con fuentes gamma que producen retrocesos electrónicos, y que el factor de eficiencia relativa de centelleo (o factor de *quenching*) de un retroceso nuclear frente a un retroceso electrónico es ≈30 % para el Na y ≈10 % para el I, la región de interés (ROI) del experimento se reduce al rango de energías menores que 10 keV, por lo que este será el intervalo en el que se va a trabajar. Cabe mencionar que, para energías tan bajas, los fotoelectrones producidos llegan de forma discreta, generando picos claramente distinguibles en la señal de los PMTs, tal y como se muestra en la figura 1. El principal problema del experimento es que, en esta región, el fondo está dominado por sucesos que no provienen del centelleo de los cristales. Estos son debidos a otras producciones de luz dentro del detector, como la luz de Cherenkov generada en los PMTs.

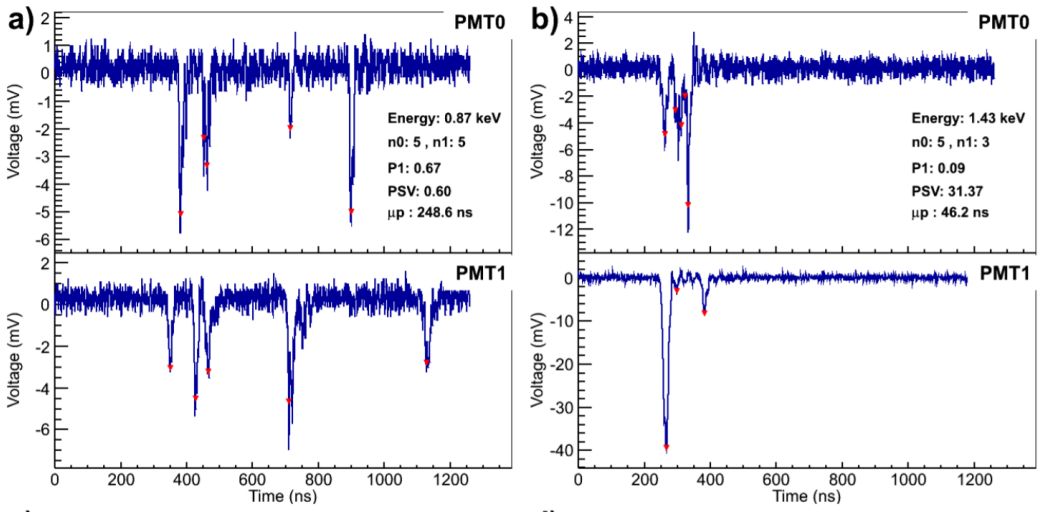

Figura 1: pulsos generados por sucesos de baja energía en ambos PMTs (0 y 1) en los que se puede apreciar con gran claridad la llegada de los phe individualmente (señalados en rojo). La imagen a) corresponde a un pulso de centelleo y la b) a un pulso con otro origen. También se añaden parámetros característicos del suceso: energía, número de picos en cada PMT (n0 y n1), P1 (relacionado con el área del pulso), primer momento (μ_p) y parámetro PSVar (PSV, definido a partir de P1 y μ_p). La figura ha sido extraída del artículo [3].

Previamente al análisis de los datos recogidos por el detector es necesario establecer una calibración en energía del mismo. Los nueve módulos son calibrados (simultáneamente) mediante fuentes externas de ^{109}Cd, proporcionando tres líneas en 11.9, 22.6 y 88.0 keV. Para que la calibración en el rango de baja energía sea más fiable, se aprovechan dos picos de energías 0.9 y 3.2 keV debidos a una pequeña contaminación de ^{22}Na y ^{40}K, respectivamente, en los cristales de NaI(Tl). Para seleccionar estos sucesos se utiliza la detección en coincidencia de la emisión de rayos X de baja energía y de radiación γ de alta energía que aparecen de forma simultánea en los procesos de captura electrónica de estos elementos. De esta forma, no solo se dispone de una mejor calibración en la ROI, sino también de una población de sucesos de centelleo a muy baja energía.

La presencia de sucesos que no tienen su origen en el centelleo de los detectores impone la necesidad de establecer unas herramientas de discriminación para eliminarlos. El método actual del experimento se basa en la definición de unos parámetros basados en la forma del pulso generado y el reparto desigual de luz entre los dos PMTs de cada módulo. Para el primer criterio se analizan: la fracción de área de la cola del pulso (100-600 ns) entre el área total (0-600 ns) (denominado P1, ver en ecuación 2) y el tiempo promedio de llegada de los fotoelectrones o primer momento (FM, ecuación 4). Estas dos variables están correlacionadas. Se asume que siguen una distribución gaussiana bidimensional y se define el parámetro PSVar de forma que un valor constante del mismo describe una elipse en el plano (P1, FM) en torno al centro de la gaussiana. La selección se hace manteniendo aquellos sucesos que posean PSVar<3, lo cual supone un porcentaje de aceptación de sucesos de centelleo (eficiencia) del 77,7 % entre 1 y 2 keV. Por otro lado, en las medidas del fondo aparecen entre 1 y 2 keV sucesos con una elevada asimetría en el número de picos detectados en cada PMT (n0 y n1). Con el objetivo de eliminarlos se establece que un suceso presente más de 4 picos en cada PMT (n0,n1>4). La eficiencia de estos cortes se establece por separado. Para el caso de PSVar se utiliza la población de coincidencia del ^{40}K y el ^{22}Na para bajas energías y simulaciones de Monte Carlo. Se les aplica el corte y se calcula la eficiencia como el número de eventos que lo satisfacen dividido entre el número total de eventos. Para el corte de n0,n1>4 se opera de forma similar, pero se trabaja sobre la población de calibración. Además de la eficiencia asociada a estos cortes, en el experimento también se tiene en cuenta la eficiencia de trigger, o probabilidad de que un suceso de una determinada energía "dispare" la electrónica de adquisición y sea registrado. En general, es muy elevada, prácticamente 1. Sin embargo, a baja energía (en torno a 1 keV), la eficiencia de detección disminuye, pues se generan pocos fotoelectrones, aumentando la diferencia promedio en los tiempos de llegada al fotocátodo. Esto provoca que la ventana temporal establecida para la coincidencia entre ambos PMT no sea lo suficientemente amplia y se pierda el suceso. En este trabajo, la eficiencia de trigger se aplicará sólo a aquellos resultados presentados como concluyentes, no a los que correspondan a análisis intermedios, aunque su efecto será poco apreciable. La eficiencia total se calcula como el producto de las tres componentes.

El espectro energético de fondo se modela mediante simulaciones de Monte Carlo realizadas con la herramienta Geant4, que permite reproducir los depósitos energéticos esperados en el detector a partir de las contaminaciones radiactivas conocidas de los distintos materiales que componen el experimento [10] . Para cada uno de los módulos se establece la simulación correspondiente y se

analiza cómo de bien se adaptan las medidas realizadas. Para valores altos de energía, desde 0,1 a 2 MeV, las simulaciones reproducen correctamente los datos experimentales, manteniendo una desviación en promedio de los 9 módulos de un 5,6 %. Para valores bajos, entre 2 y 6 keV, esta desviación sigue siendo pequeña, con un valor de -7,9 % (ahora con valor negativo por ser la medida experimental superior a la simulación). Sin embargo, el problema surge en el intervalo de 1 a 2 keV, como se puede ver en la figura 2, donde la desviación aumenta hasta el -48 %.

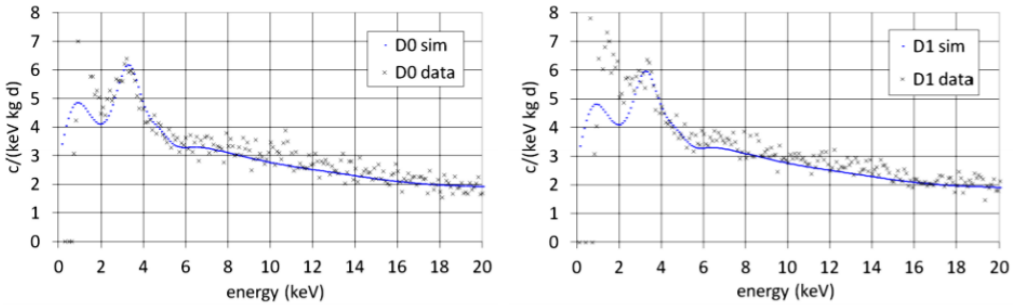

Figura 2: resultados de los ritmos medidos en los detectores D0 y D1 (data) junto a los valores extraídos mediante simulaciones (sim). Se puede apreciar la diferencia que aparece por debajo de 2 keV. Figuras extraídas del artículo [10].

Este intervalo es en el que interesa trabajar. Esta discrepancia puede ser debida a sucesos de ruido que no han sido filtrados mediante los métodos mencionados o a la falta de una componente en el modelo de fondo que no ha sido considerada en la simulación.

Además de los 9 módulos que componen el experimento ANAIS-112, se dispone de un módulo más idéntico al resto, a excepción de que no posee un cristal centelleador en su interior (módulo blank). Este es utilizado para analizar los sucesos producidos dentro de los propios fotomultiplicadores, y que, por tanto, representan una población de sucesos "no de centelleo" que se utilizará en la fase de entrenamiento del algoritmo.

3. Diseño del algoritmo

El *Boosted Decision Tree* (BDT) es un algoritmo basado en la generación de varios clasificadores, denominados árboles, que se aplican sobre los diferentes sucesos de forma combinada. Así se consigue establecer un criterio de selección más adecuado y preciso que el que podría generar un solo árbol. El entrenamiento del algoritmo opera con dos poblaciones: una que engloba sucesos de señal ("buenos") y otra correspondiente a sucesos de ruido[1] ("malos"). Además de escoger dichas poblaciones se establecen las variables con las que se quiere llevar a cabo la clasificación. Estas son seleccionadas en base a las diferencias conocidas entre los sucesos de señal y de ruido, carece de sentido tratar de separarlos empleando una variable cuyo valor sea muy similar para ambas poblaciones.

[1] Aunque la terminología convencional los denomina sucesos de fondo [11], en este trabajo se ha escogido esta terminología para evitar confusiones con las medidas del experimento.

Cada uno de los árboles se genera según el siguiente criterio de construcción, cuyo esquema se muestra en la figura 3. Se parte de un nodo base en el que se introducen los sucesos sin clasificar y que, en general, contendrá una mezcla de ambas poblaciones (señal, s y ruido, r). Cada evento lleva asociado un peso $\omega^{s,r}$ y sus valores para las diferentes variables de entrenamiento definidas. Normalmente, los pesos son establecidos de forma que, en el nodo inicial, la suma de los correspondientes al ruido y a la señal coincidan. En este caso, se trabaja con el mismo número de sucesos de señal que de ruido por lo que se les asigna a todos el mismo peso. A continuación, se recorren las diferentes variables de entrenamiento, buscando el valor de corte que mejor separa las dos poblaciones, y se selecciona aquella variable cuyo corte establece la mejor división. De esta forma, los eventos del nodo son dirigidos a otros dos nodos, según si superan o no el corte. En estos nuevos nodos se vuelve a aplicar la misma operación y se realiza de forma iterativa hasta llegar a un nodo en el cual se cumple un cierto criterio de detención. Dicho nodo se denomina "hoja" (muy acorde al término árbol). Cabe mencionar que existe la posibilidad de que en un mismo árbol se repitan variables en diferentes nodos.

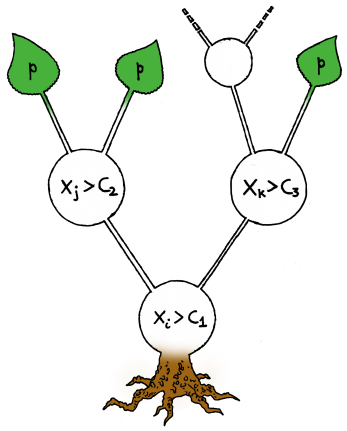

Figura 3: esquema de un árbol que trabaja con un conjunto de variables $(x_1,\ x_2, ..., x_n)$, estableciendo cortes $(c_1,\ c_2, ..., c_m)$. Algunos caminos finalizan en hojas caracterizadas por su valor de pureza (p).

Una vez construido el árbol, se le introduce el conjunto de eventos que se quiere clasificar. Dado uno de ellos (i), se le va aplicando el criterio de corte de cada uno de los nodos, siguiendo una u otra dirección, hasta que acaba en una de las hojas. Entonces, se le asigna el valor asociado a esta hoja. Este valor puede ser definido de varias maneras, por ejemplo, puede corresponder a la pureza del nodo: $p = s/(s + r)$, siendo s la suma de los pesos de señal y r la de ruido que han finalizado en esa hoja durante el entrenamiento. Este valor caracteriza a la hoja y se le asigna una etiqueta según la predominancia de los eventos que la ocupan. Si p$\geq$0.5, a la hoja se le asigna un valor +1 y la etiqueta de *señal*, mientras que, si p$<$0.5, se le asigna -1 y la etiqueta de *ruido*.

Trabajar con uno solo de estos árboles no suele ser suficiente para obtener una buena clasificación. Aquí es donde entra en juego el *Boost* del algoritmo. En términos generales, este proceso consiste en la aplicación de varios árboles a los diferentes sucesos, formando un "bosque". Cada vez que son clasificados se les asigna una modificación en su peso. Esta se basa en la clasificación producida en el árbol de forma que el peso de un evento clasificado correctamente, es decir, que ha acabado en una hoja cuya etiqueta coincide con su procedencia, no cambia, mientras que uno clasificado incorrectamente sufrirá un incremento en su peso. Así, en el siguiente árbol se introducen los sucesos con sus pesos modificados. Aquellos con mayor peso implican una mayor relevancia en el proceso de clasificación de los eventos, es decir, el siguiente árbol buscará criterios de corte centra-

dos en estos eventos. Los árboles iniciales engloban los cortes más gruesos, realizando la separación principal entre sucesos, mientras que la siguiente concatenación de árboles puede interpretarse como una serie de correcciones más precisas a dichos cortes. El valor que se le asignará a un evento clasificado por este conjunto de árboles dependerá de su paso a través de todos ellos, combinando los valores asociados a las hojas de cada árbol en el que han acabado. Así, se tiene el mismo conjunto de sucesos, pero con una nueva variable asociada denominada, de forma general, BDT (ver ecuación 1 [4]).

$$BDT(\vec{x}_i) = \frac{1}{nTrees} \cdot \sum_{j=0}^{nTrees} [\ln(\alpha_j) \cdot T_j(\vec{x}_i)] \tag{1}$$

donde nTrees es el número de árboles, $\alpha_j = (1 - f_j)/f_j$, siendo f_j la fracción de sucesos clasificados incorrectamente y $T_j(\vec{x}_i)$ el valor de pureza del suceso correspondiente al árbol j-ésimo. Una vez ha sido asignada, se aplican cortes sobre esta que eliminan aquellos sucesos con un valor inferior.

El algoritmo con el que se va a trabajar está compuesto por dos fases, en cada una de ellas se realizará la clasificación de sucesos en base a variables que caracterizan diferentes aspectos de los mismos. Sin embargo, ambas mantendrán un conjunto de parámetros comunes en la construcción de los árboles (y del bosque). Entre ellos resulta interesante remarcar los siguientes:

- nTrees: corresponde al número de árboles que componen el algoritmo.
- MinNodeSize=2,5 %: establece el porcentaje mínimo de sucesos de entrenamiento que debe tener un nodo hoja.
- MaxDepth=3: indicando la profundidad máxima que puede alcanzar un árbol. Este valor limita el número de divisiones que puede realizar una muestra partiendo desde el nodo raíz.
- BoostType=AdaBoost: selecciona el algoritmo de *Boosting* a implementar. El *Adaptive Boost* es el más común, destacado por su capacidad de adaptar el proceso de entrenamiento a los datos con los que se entrena. Como se establece en la sección 7.1 del manual [4], trabaja bien con árboles de poca profundidad, de acuerdo con el valor que se ha establecido en este programa.
- SeparationType=GiniIndex: corresponde al criterio de división de nodos. Este se basa en la generación de una función, *Gini Index*, relacionada con la pureza de los nodos. Se busca entonces el corte en cada nodo que minimice la impureza del mismo en base a los valores de esta función. Como menciona el artículo [11], este método es el más utilizado en este tipo de programas.
- NCuts=20: limita el número de cortes buscados para separar el nodo. Este valor es indicado por el manual [4] como un buen compromiso entre eficiencia y tiempo de computación.

A excepción de nTrees, estos parámetros permanecerán constantes en el algoritmo.

3.1. Primera Fase: análisis de la forma del pulso

En esta fase, el entrenamiento y la clasificación se va a basar en la forma de los pulsos generados en cada uno de los sucesos detectados.

Los sucesos de calibración (^{109}Cd) serán la población de señal y los del módulo blank, la de ruido. Concretamente, la calibración engloba todos los datos obtenidos durante 3 años y las del módulo blank corresponden a los tomados en los años 2 y 3. De estas poblaciones se van a utilizar los sucesos comprendidos entre 1 y 2 keV, dado que es el rango en el que predominan los eventos de ruido. Las variables empleadas para la clasificación son aquellas que se relacionan con la forma del pulso:

- P1: de forma cualitativa se puede entender como la fracción del área del pulso correspondiente a la cola del mismo (ecuación 2). Puesto que los sucesos del PMT son mucho más rápidos que los de centelleo, su valor de P1 será mucho menor.
- P2: similar a la variable P1, pero refiriéndose a la fracción de la cabeza del pulso (ecuación 3).
- FM: el primer momento o tiempo medio, calculado desde la posición establecida del trigger (t_0) hasta el final del pulso (t_f) sobre el número de picos identificados (ecuación 4).
- CAPx: la función distribución del pulso, que define la fracción de área del pulso comprendida entre 0 ns y el valor x ns establecido respecto al área total, desde t_0 a t_f (ecuación 5). Se aplicará para x=50, 100, 200, 300, 400, 500, 600, 700, 800.

$$P1 = \frac{\sum_{100\ ns}^{600\ ns} A_i}{\sum_{0\ ns}^{600\ ns} A_i} \qquad (2)$$

$$P2 = \frac{\sum_{0\ ns}^{50\ ns} A_i}{\sum_{0\ ns}^{600\ ns} A_i} \qquad (3)$$

$$FM = \frac{\sum_p A_p \cdot t_p}{\sum_p A_p} \qquad (4)$$

$$CAPx = \frac{\sum_{0\ ns}^{x\ ns} A_i}{\sum_{0\ ns}^{t_f - t_0\ ns} A_i} \qquad (5)$$

el elemento A_i corresponde al área del pulso en la ventana temporal t_i. De forma similar, A_p y t_p son el área y el tiempo de cada pico (p) identificado en el pulso (ver figura 1).

Al aplicar esta fase a un conjunto de sucesos, el algoritmo asignará a cada uno la variable BDT, con el valor obtenido de la clasificación.

3.2. Segunda fase: análisis de la asimetría del pulso

Un suceso de centelleo produce una distribución isótropa de luz, por lo que ambos fotomultiplicadores registrarán pulsos de intensidad similar. Por otro lado, un proceso que genere luz en uno de los PMT producirá más intensidad en dicho PMT que en el otro. Como ya se ha introducido, por debajo de 2 keV dominan los sucesos que presentan una clara asimetría respecto al reparto de luz entre los dos fotomultiplicadores. Mientras que en uno de los PMT se genera un número bajo de picos de voltaje, en el otro este número es elevado. Dichos sucesos no aparecen en las medidas

de calibración de ^{109}Cd, por ello, esta población será utilizada como la de señal en el entrenamiento. La población de ruido corresponderá a los eventos en los que aparece esta asimetría, es decir, los medidos en el fondo[2]. Al igual que en la fase anterior, se restringen al rango de 1 a 2 keV. A pesar de que el fondo no es una población pura de sucesos de ruido, en el rango utilizado estos son los dominantes, por ello puede ser utilizada en el entrenamiento como población de ruido. Para incorporar la población de calibración al entrenamiento se aprovechará la fase anterior. Dado que la población de calibración puede tener una pequeña contaminación de sucesos malos, se establece un corte orientativo de BDT que elimine parte de estos. Así, se introduce una población de señal más pura en el entrenamiento.

Las variables iniciales escogidas para entrenar el algoritmo en esta fase son:

- n0, n1: el número de picos detectados en el fotomultiplicador 0 y 1, respectivamente.
- Asyarea: esta variable define la asimetría entre el área del pulso medido en cada uno de los PMT, $area0$ y $area1$ (calculados como el área del pulso integrada en toda su traza, desde t_0 hasta t_f).

$$Asyarea = \frac{area0 - area1}{area0 + area1} - < A > \tag{6}$$

  Dado que los PMT acoplados a cada centelleador poseen diferente ganancia, siempre va a aparecer un cierto grado de asimetría entre sus pulsos. El término $< A >$ corresponde al valor medio de esta asimetría, calculado a partir de la calibración con ^{109}Cd, para cada detector y bin de 1 keV. Al sustraerlo, se fija que $Asyarea$ se refiera solo a asimetrías debidas a motivos externos a la ganancia de los fotomultiplicadores y permite combinar los datos de los diferentes detectores.

En esta fase, la variable que se le añade a los sucesos tras aplicarles el bosque entrenado se denomina BDT2.

3.3. Eficiencia

La eficiencia del método incorporado se calcula sobre la calibración utilizada como población de señal en el entrenamiento de ambos conjuntos de variables. Para cada intervalo de energía se aplicará el corte de BDT (y BDT2 en la segunda fase) establecido y se calculará el cociente entre el número de eventos seleccionados y la población total. Así, la eficiencia consiste en la fracción de eventos de calibración que sobreviven al corte. Esta eficiencia será utilizada para corregir el ritmo de fondo tras aplicar los cortes establecidos. Esta corrección consiste en dividir el valor del ritmo en cada bin de energía entre el valor de la eficiencia de dicho bin. Al igual que para las eficiencias dadas por los cortes de PSVar, n0 y n1, se introducirá el factor debido a la eficiencia de trigger en los resultados correspondientes (como se ha indicado en la introducción).

[2]No se hace uso de la población del módulo blank porque en este no es posible definir el término $< A >$ (ver ecuación 6).

3.4. Proceso de optimización del algoritmo

Se va a trabajar buscando modificaciones en los parámetros de entrenamiento de ambas fases que proporcionen una mejor discriminación de los sucesos de ruido entre 1 y 2 keV de la PT manteniendo un compromiso con la eficiencia de detección y el tiempo de computación. Por ello, lo que se busca en cada etapa de este proceso son los valores de corte de BDT y BDT2 que minimicen el nivel de fondo manteniendo una eficiencia igual o superior a la del análisis estándar de ANAIS-112. Se va a trabajar con:

1. Número de árboles: se van a construir varios bosques, variando el número de árboles que los componen. Se analizará el tiempo de computación necesario para ello y el número mínimo de árboles que permiten una reducción del ritmo apreciable.

2. Entrenamiento de los detectores por separado: de base, el programa trabaja con las poblaciones de sucesos del conjunto de los nueve detectores. Se va a estudiar cómo varía el resultado al entrenar cada detector solo con los sucesos del mismo.

3. Expansión en el número de variables: se incorporarán más variables al entrenamiento. Estas estarán relacionadas con aspectos de asimetría de los pulsos por lo que se aplicará a la segunda fase.

4. Combinar ambas fases: el criterio de utilizar dos fases se basa en la imposibilidad de definir la variable Asyarea en el módulo blank. Sus medidas resultan ideales para realizar la primera fase del programa, pero no sirven para la discriminación en aspectos de asimetría, por lo que se requiere la segunda fase. No obstante, se va a intentar trabajar con un solo entrenamiento, aplicando todas las variables y usando las medidas de fondo del experimento como la población de ruido.

4. Incorporación de las modificaciones

4.1. Número de árboles del entrenamiento

La selección del número de árboles óptimo (*nTrees*) se va a realizar para ambas fases. Tras analizar la primera se fijará el valor que se considere más adecuado y se utilizará para preparar los datos para la siguiente fase.

4.1.1. Primera Fase

Interesa recorrer un amplio rango de valores, para ver cómo se comporta el algoritmo en ambos extremos: muchos y pocos árboles. En la figura 4 se muestra el tiempo de computación empleado para los entrenamientos realizados con 10, 50, 100, 500, 850, 1000 y 2000 árboles. Como se puede intuir, este tiempo será mayor de acuerdo al aumento de los árboles entrenados. Se observa que sigue una dependencia lineal.

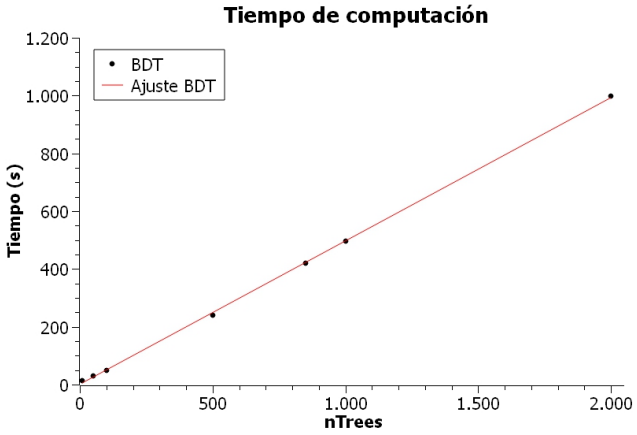

Figura 4: representación gráfica del comportamiento del tiempo de entrenamiento frente al número de árboles involucrado. En rojo se presenta un ajuste lineal realizado en *SciDavis* que describe adecuadamente la tendencia observada..

El tiempo de computación es una variable importante en la optimización de un algoritmo por lo que esta información resultará relevante más adelante.

Para analizar los resultados del entrenamiento, resulta interesante comparar cómo se distribuye la variable BDT en el conjunto de sucesos de entrenamiento (calibración y módulo blank), así como en la población objetivo (PT). La figura 5 muestra los histogramas de la variable BDT para sucesos de calibración, módulo blank (primera y segunda columna respectivamente) y fondo del detector D0 (tercera columna) entre 1 y 2 keV (línea roja) para tres valores concretos de número de árboles (10, 850 y 1000, uno por fila). Sobre ellos se representa la selección de eventos con P1<0,2 (verde) y P1>0,4 (azul), lo que permite hacerse una idea de cómo se distribuyen los sucesos "malos" (con valores pequeños de P1) y dónde los "buenos" (mayor P1).

Hay varios aspectos a destacar. En primer lugar, se puede observar cómo los sucesos de P1 bajo aparecen en los valores más pequeños de BDT (sucesos "malos"), mientras que los de P1 alto, aunque ocupan mayor rango, tienden a orientarse hacia BDT mayores. Como cabía esperar, en el caso de la calibración, la mayor parte de los sucesos aparecen en la zona de BDT alto y en el caso del blank, en el BDT bajo. En la distribución de los sucesos de fondo se puede observar cómo los sucesos "malos" son los más abundantes. Finalmente, se puede apreciar cómo cambian las distribuciones al aumentar el número de árboles. Para un número bajo (10) la distribución resulta más discreta, mientras que para bosques más grandes el proceso de clasificación es más preciso y permite una mayor continuidad del espectro. Por otro lado, el resultado para 850 y 2000 árboles es muy similar (aunque cambie el rango de BDT, que resulta poco relevante si la distribución coincide) lo que da una primera idea de que posiblemente no sea necesario trabajar con tantos árboles.

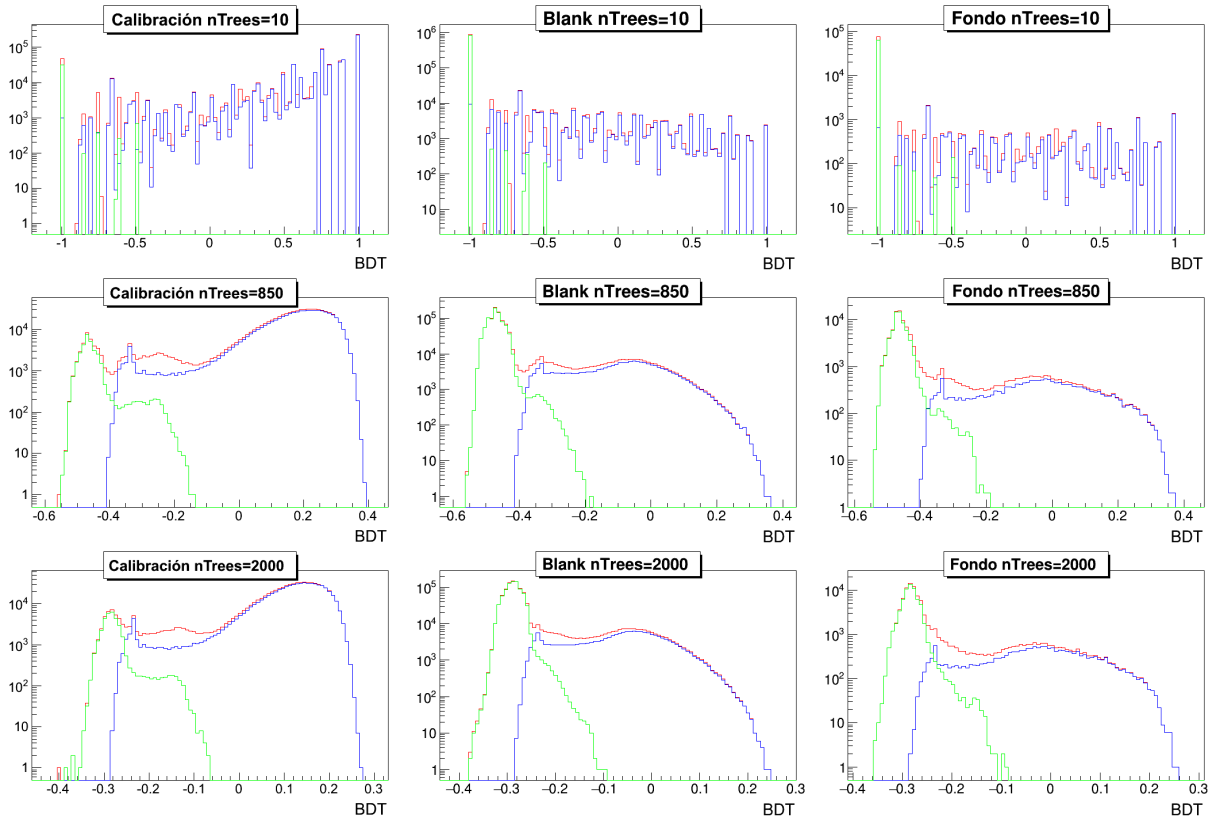

Figura 5: histogramas de la variable BDT seleccionados para 10, 850 y 2000 árboles de las poblaciones de calibración, blank y fondo del detector D0 (rojo). En azul se muestra la selección con P1>0,4 y en verde con P1<0,2.

El siguiente paso es encontrar el valor de corte de BDT que permita reducir el ritmo en el intervalo de 1 a 2 keV eliminando, preferentemente, sucesos malos. No obstante, dado que las variables definidas sobre las poblaciones de entrenamiento no presentan una separación total en ningún caso, es posible que varios eventos buenos posean valores de BDT similares a los malos, lo que provoca que el corte de BDT escogido los rechace. Para corregir este efecto, se calculará la eficiencia del corte como se ha explicado en la sección 3.3 y se representará el ritmo una vez corregido por esta. Para evitar cortes demasiado agresivos que eliminen la mayor parte de los sucesos, se va a trabajar con la eficiencia como variable de control, imponiendo que esta no sea inferior a la eficiencia del análisis estándar de ANAIS-112 [3] (PSVar en esta primera fase).

La figura 6 representa el ritmo medio entre 1 y 2 keV para diferentes valores de corte de BDT. Por mantener la elección previa, se utiliza el detector D0 para la prueba. Como parámetro de control se escoge la eficiencia media entre 1 y 10 keV, cuyo valor dado por el corte de PSVar es del 96,3 %. El ritmo no comienza a decrecer hasta que el corte de BDT empieza a eliminar sucesos. Como se muestra en la figura 5, esto sucede para valores más grandes conforme mayor número de árboles tenemos. Lo que se concluye de este análisis es comprobar que el ritmo mínimo obtenido para 500 o menos árboles es notoriamente superior, por lo que se considerarán solo los conjuntos de 850, 1000 y 2000 árboles.

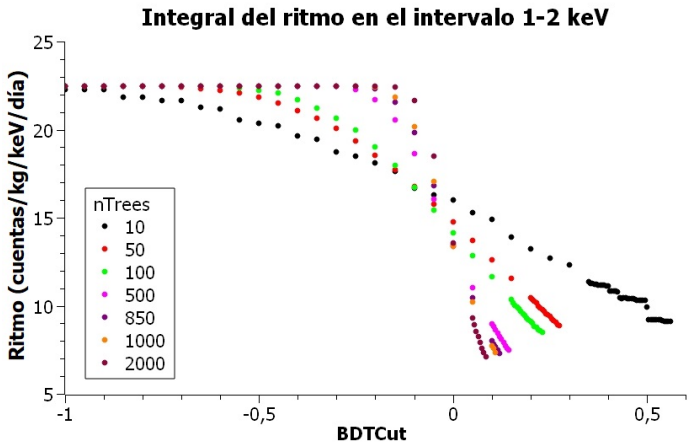

Figura 6: representación del ritmo medio del detector D0 entre 1 y 2 keV para diferentes valores de corte de BDT, limitados por una eficiencia media entre 1 y 10 keV de 96,3 %.

En la figura 7 se muestra el caso de 850 con el corte correspondiente (BDTCut=0.120). Se puede comprobar cómo la eficiencia entre 1 y 2 keV es superior a la de PSVar en todos los bines de energía, compensando las pequeñas variaciones que hay en energías mayores, igualmente observable para 1000 y 2000 árboles (presentados en el Anexo I). Ante estos resultados, se decide que la eficiencia media entre 1 y 10 keV supone un buen parámetro de control.

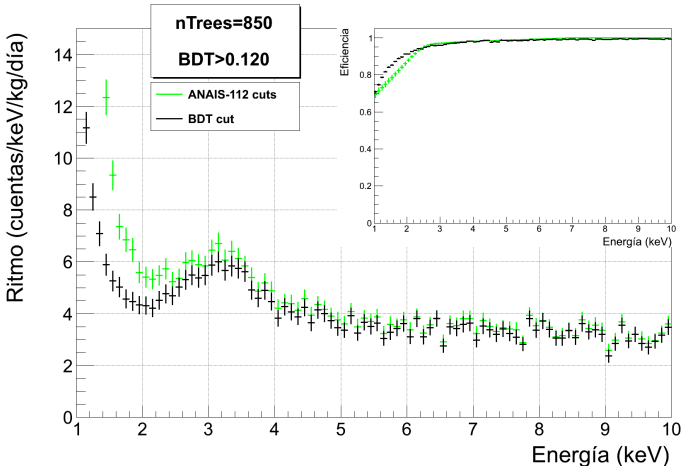

Figura 7: comparación del ritmo en el detector D0 en función de la energía tras aplicar el filtrado y la corrección por la eficiencia con el análisis estándar (en verde) y el análisis BDT con 850 árboles y BDT>0.120 (en negro). En el recuadro se representa la eficiencia para ambos casos.

Aunque, en una primera aproximación se ha estado aplicando el método de corte solo para el detector D0, lo que interesa es establecerlo para el conjunto de los nueve detectores. Para ello, los siguientes estudios se van a efectuar utilizando para cada parámetro el promedio de los nueve detectores a los que se les aplica el mismo corte. Se realiza un análisis más preciso para extraer el valor de corte, los resultados obtenidos se presentan en la tabla 1.

nTrees	BDTCut	Ritmo 1-2 keV (cuentas/kg/keV/día)	Eff.media
850	0,106	6,41±0,05	0,9635±0,0007
1000	0,099	6,38±0,05	0,9636±0,0007
2000	0,074	6,26±0,05	0,9638±0,0007

Tabla 1: valores de corte resultantes de la primera fase del entrenamiento para el conjunto de los 9 detectores, promediando su eficiencia entre 1 y 10 keV y el ritmo entre 1 y 2 keV.

Los valores del ritmo mínimo para 850 y 1000 árboles son compatibles entre sí teniendo en cuenta su incertidumbre, la cual no sufre grandes modificaciones entre los diferentes conjuntos. Aunque el ritmo obtenido con 2000 árboles es ligeramente inferior, el tiempo de computación es muy superior a los otros dos casos pues, como se ha mostrado anteriormente, es directamente proporcional al número de árboles. Por consiguiente, se establece 850 como el valor adecuado para esta primera fase.

Esto implica que en la segunda fase se trabajará con las poblaciones de sucesos que hayan sido previamente analizados con los 850 árboles y posean ya el valor de BDT correspondiente.

4.1.2. Segunda Fase

El proceso va a ser parecido al de la fase anterior, buscando entre 10, 50, 100, 850, 1000 y 2000 árboles el valor óptimo. Las variables utilizadas en el entrenamiento son menos que para el análisis de la forma del pulso (3 frente a 12), por lo que se espera que el número de árboles necesarios sea también menor.

Tras el entrenamiento de la segunda fase, la distribución de la variable asignada BDT2 para la PT del detector D0 se muestra en la figura 8 (línea roja). En la misma figura se representa la selección de de sucesos del análisis estándar de ANAIS-112, PSVar<3 y n0,n1 mayor que 4 (azul).

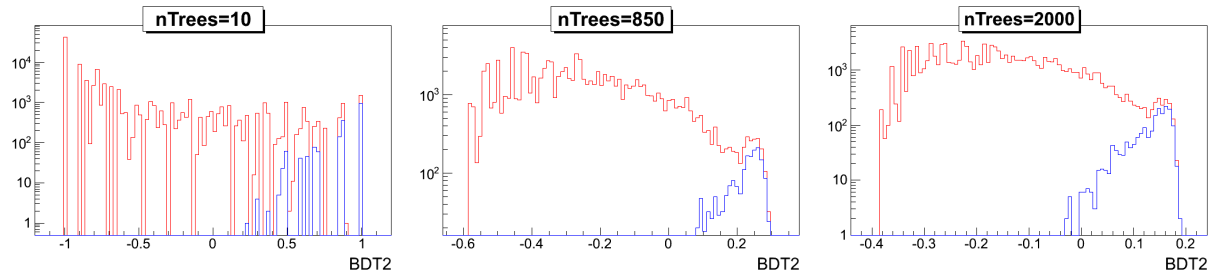

Figura 8: distribución de la variable BDT2 de las medidas de fondo del detector D0 para el intervalo de energía entre 1 y 2 keV habiendo trabajado con diferentes números de árboles. En azul se superpone la selección para PSVar<3 y n0,n1>4.

Como ya se ha mencionado, entre 1 y 2 keV el fondo sufre una gran contaminación de eventos no provenientes del centelleo del NaI(Tl). Esto se relaciona con que la mayor concentración de sucesos se da para valores bajos de BDT2, es decir, que son "malos". Por otro lado, aquellos seleccionados por los cortes de ANAIS tienden a agruparse en valores altos. Lo más interesante de estas repre-

sentaciones es que muestran como, con tan solo 10 árboles, ya se puede conseguir una correcta separación entre sucesos.

En la fase anterior se trabajaba en un campo bidimensional dado por el corte de BDT y el ritmo entre 1 y 2 keV. En este caso tenemos tres dimensiones: las dos anteriores y el corte de BDT2. Entonces, para realizar la búsqueda de los dos valores de corte que optimicen el resultado se recurre a un análisis de rejilla. Para cada valor de corte de BDT se recorren los diferentes valores de corte de BDT2, hasta que la eficiencia media entre 1 y 10 keV es igual al valor dado por los cortes realizados por ANAIS. Este es 94,06 % (promediado para los 9 detectores).

Este tipo de estudio requiere gran tiempo de computación por lo que se decide aplicarlo solo a las poblaciones de 10, 850 y 2000 árboles, de forma que se pueda formar una visión general del resultado en base a los dos extremos (10, 2000) y el punto intermedio (850). La tabla 2 muestra los resultados obtenidos mediante este método. Para los tres conjuntos se consiguen unos valores del ritmo compatibles estadísticamente. Tal y como se había planteado en base a la figura 8, el uso de 10 árboles ya resulta suficiente en esta segunda fase. Esto supone una gran ventaja, pues reduce en gran medida el tiempo de computación necesario (frente a necesitar 850 árboles como en la fase anterior, ver Anexo II).

nTrees	**BDTCut**	**BDT2Cut**	**Ritmo 1-2 keV (cuentas/kg/keV/día)**	**Eff.media**
10	0,10	0,69	4,11±0,05	0,9435±0,0007
850	0,10	0,14	4,13±0,06	0,9421±0,0007
2000	0,11	0,07	4,14±0,05	0,9421±0,0007

Tabla 2: análisis realizado tras aplicar las dos fases del entrenamiento para el conjunto de los 9 detectores, promediando su eficiencia entre 1 y 10 keV y el ritmo entre 1 y 2 keV.

4.1.3. Resultados

El estudio realizado conduce al establecimiento del uso de 850 árboles para la primer fase y 10 para la segunda. De esta forma, se obtiene el ritmo en el intervalo de 1 a 2 keV presentado en la tabla 3.

BDTCut	**BDT2Cut**	**Ritmo 1-2 keV (cuentas/kg/keV/día)**	**Eff.media**
0,10	0,69	4,11±0,05	0,9425±0,0007

Tabla 3: resultados finales tras el análisis del número de árboles en ambas fases.

En estos valores, como se ha mencionado en la introducción, viene incorporado el factor de la eficiencia de trigger. En la figura 9 se puede observar la comparación del ritmo y la eficiencia para ambos métodos. Además, en la tabla 4 se puede ver una rápida comparación con los valores dados al aplicar el método de ANAIS, mostrando la eficacia del algoritmo utilizado. Se puede comprobar cómo para el intervalo entre 1 y 2 keV el ritmo se reduce cerca de un 30 % mediante el método de BDT y BDT2 respecto al dado por ANAIS, mientras que, entre 2 y 10 keV, apenas se modifica

(lo cual era esperable, ya que en este rango energético la simulación Monte Carlo proporciona una buena descripción de la medida experimental y no tenemos motivos para sospechar que haya contaminación de ruido).

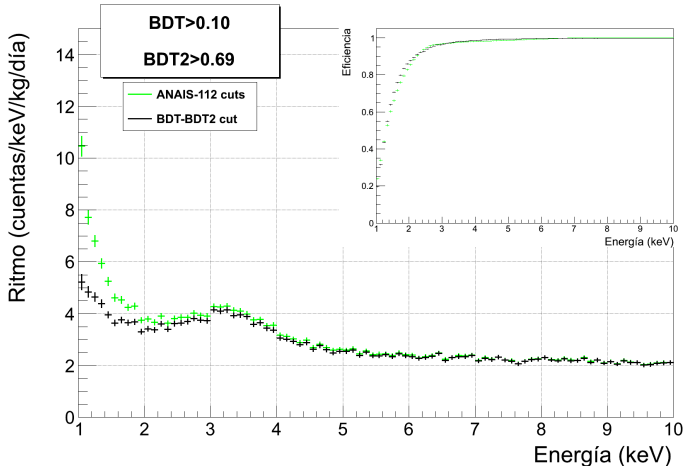

Figura 9: representación del ritmo y de la eficiencia en el intervalo de 1 a 10 keV aplicando los cortes de ANAIS (en verde) y los de BDT-BDT2 (en negro), habiendo trabajado con 850 (primera fase) y 10 (segunda fase) árboles.

Diferencia del ritmo 1-2 keV	**Diferencia del ritmo 2-10 keV**
28,6 %	2,7 %

Tabla 4: diferencia entre el ritmo promedio obtenido mediante los cortes de ANAIS y los del método BDT y BDT2.

4.2. Entrenamiento de cada detector por separado

En la sección 4.1, los entrenamientos se han llevado a cabo utilizando conjuntamente los sucesos recogidos por los 9 detectores (y por el módulo blank). A continuación, se entrenará cada uno de los detectores por separado, utilizando únicamente los sucesos de dicho detector. Por un lado, esto puede ayudar a discernir posibles diferencias entre los detectores, pero, por otro, puede que la mayor variedad de sucesos en las poblaciones de entrenamiento resulte favorable o, realmente, no sea suficiente como para producir un efecto apreciable.

Aprovechando los resultados del apartado anterior, se va a trabajar con 850 árboles en la primera fase y 10 en la segunda para cada uno de los detectores. Para el entrenamiento se debe dividir la población de calibración y la de ruido de la segunda fase según el detector. En el entrenamiento se seguirán utilizando el mismo número de sucesos de señal y de ruido. Una vez se han entrenado todos los bosques y se han aplicado a las medidas de la PT de los detectores correspondientes se presenta en la figura 10 cómo cambia su distribución frente a BDT y BDT2.

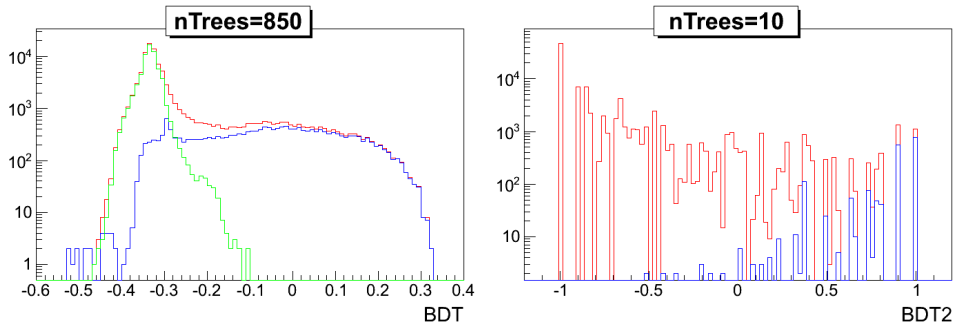

Figura 10: distribución de las variables BDT (izquierda) y BDT2 (derecha) de los sucesos de fondo medidos por el detector D0 tras haber entrenado el algoritmo con sucesos provenientes del mismo. En el histograma de BDT se han seleccionado los eventos con P1<0,2 (verde) y P1>0,4 (azul) y en el de BDT2, los de PSVar<3 y n0,n1>4 (azul).

Comparando con las figuras 5 y 8 se puede apreciar cierto cambio, especialmente respecto a BDT2, en la que los sucesos con PSVar<3 y n0,n1>4 están más ampliamente distribuidos que en el caso anterior. Esto no implica necesariamente una peor distribución, pues es posible que los sucesos que ahora aparecen en valores de BDT2 menores sean eliminados al establecer cortes relacionados con la forma del pulso (BDT).

Se procede a realizar el mismo análisis en rejilla utilizado en la sección 4.1.2 para obtener los valores de corte de BDT y BDT2. El parámetro de control, así como su valor límite, siguen siendo los mismos (eficiencia entre 1 y 10 keV $\geq$94,06 %).

4.2.1. Resultados

Mediante el análisis realizado se obtienen los valores presentados en la tabla 5. El valor del ritmo entre 1 y 2 keV ha sido reducido, pero no de una forma relevante. El rango de valores dado por el intervalo de incertidumbre lo hace compatible con el resultado dado para el entrenamiento con todos los sucesos juntos. Dado que en este caso se debe entrenar y aplicar un bosque para cada detector en cada una de las fases, el tiempo de computación es mucho mayor, resultando poco rentable. Por consiguiente, se sigue trabajando con la población total de calibración en el entrenamiento.

BDTCut	**BDT2Cut**	**Ritmo 1-2 keV (cuentas/kg/keV/día)**	**Eff.media**
0,10	0,86	3,98±0,06	0,9414±0,0007

Tabla 5: análisis realizado para el conjunto de los 9 detectores, promediando su eficiencia entre 1 y 10 keV y el ritmo entre 1 y 2 keV. El algoritmo ha sido entrenado para cada detector con los sucesos detectados en cada uno de ellos.

4.3. Extensión de variables

Los sucesos poseen otras variables definidas que no han sido incorporadas en el entrenamiento y aplicación del algoritmo. Estas podrían ayudar a generar una mejor clasificación de los eventos,

por lo que van a ser añadidas para observar cuánto varían los resultados obtenidos. Dichas variables están relacionadas con la posible asimetría del suceso detectado en ambos PMTs:

- P10-P11: estas variables corresponden a la misma definición que la presentada en la ecuación 2, pero definidas por separado para cada uno de los fotomultiplicadores.
- nphe0-nphe1: identifican el número de phe y se calculan como el área del pulso del PMT correspondiente entre el área media de un phe. Este valor depende del voltaje con el que se alimenta el PMT.
- Asynphe: de forma similar a Asyarea, esta variable calcula la diferencia entre nphe0 y nphe1.

$$Asynphe = \frac{nphe0 - nphe1}{nphe0 + nphe1} \tag{7}$$

- P20-P21: son análogas a P10 y P11, pero aplicando la definición dada en la expresión 3.

El procedimiento es el mismo que se ha llevado a cabo en los apartados anteriores, analizando los casos de 10, 850 y 2000 árboles. La diferencia es que se van a ir incorporando las variables mencionadas. En concreto se prueba añadiendo cada conjunto presentado por separado a n0, n1 y Asyarea, y a añadirlas todas en el mismo entrenamiento. En la figura 11 se presentan las distribuciones que se obtienen para las medidas de fondo del detector D0 frente a BDT2 para el caso de 850 árboles (que permite visualizar más claramente los resultados que el caso de 10 y es altamente similar para el de 2000) en rojo, superponiendo en azul la selección dada por los cortes de ANAIS.

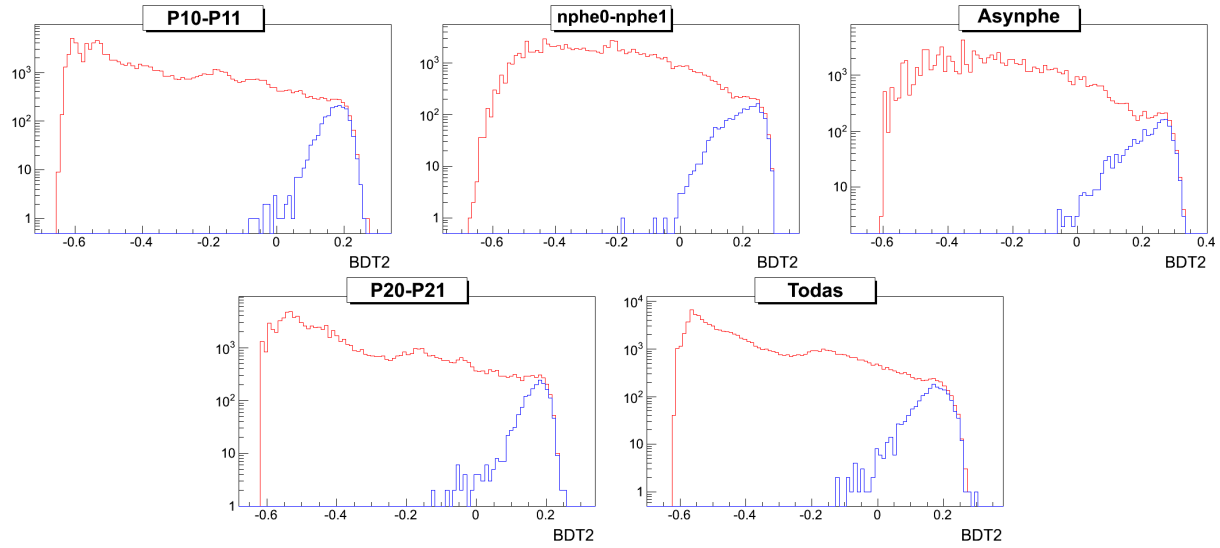

Figura 11: histogramas de la distribución de sucesos de fondo del detector D0 respecto a la variable BDT2 habiendo añadido más variables al entrenamiento (indicadas sobre cada gráfico) para el caso de 850 árboles. En azul se representa la distribución de sucesos con PSVar<3 y n0,n1>4.

Atendiendo a la distribución general de los sucesos se puede apreciar que es muy similar para los conjuntos con P10-P11, P20-P21 y añadiendo todas las variables, mientras que para los casos

de nphe0-nphe1 y Asynphe esta cambia ligeramente, agrupando más sucesos en una zona cercana al centro que en el extremo de bajos valores de BDT2. No obstante, en todos los casos se encuentra que los eventos seleccionados por ANAIS quedan agrupados en los valores más altos de BDT2.

Dado que se están añadiendo más variables al entrenamiento, ya no se puede asegurar que 10 árboles sean suficientes, por lo que se vuelve a trabajar con 10, 850 y 2000 árboles para cada conjuntos de variables añadido.

4.3.1. Resultados

Trabajando con la eficiencia media como parámetro de control y siguiendo el análisis en rejilla se obtienen los resultados de la tabla 6 para las diferentes adiciones de variables de entrenamiento. Antes de nada, cabe mencionar que para los tres conjuntos de nTrees se consiguen resultados compatibles por su intervalo de incertidumbre, a excepción de la adición de todas las variables, donde el uso de 10 árboles da el mejor resultado. Los valores obtenidos no distan demasiado de lo que ya se conseguía trabajando solo con n0, n1 y Asyarea. De hecho, sin atender a los intervalos de incertidumbre, el menor valor para el ritmo se consigue para el caso de 10 árboles en nphe0-nphe1.

	nTrees	**BDTCut**	**BDT2Cut**	**Ritmo 1-2 keV (cuentas/kg/keV/día)**	**Eff.media**
P10-P11	10	0,12	0,32	4,23±0,05	0,9397±0,0007
	850	0,12	0,09	4,17±0,05	0,9396±0,0025
	2000	0,12	0,06	4,18±0,05	0,9399±0,0007
nphe0-nphe1	10	0,10	0,68	4,10±0,06	0,9396±0,0007
	850	0,12	0,06	4,21±0,05	0,9405±0,0007
	2000	0,12	0,04	4,20±0,05	0,9405±0,0007
Asynphe	10	0,11	0,55	4,18±0,05	0,9399±0,0007
	850	0,12	0,07	4,17±0,05	0,9396±0,0007
	2000	0,12	0,03	4,25±0,05	0,9416±0,0007
P20-P21	10	0,12	0,50	4,15±0,05	0,9396±0,0007
	850	0,12	0,09	4,18±0,04	0,9415±0,0007
	2000	0,12	0,06	4,17±0,04	0,9419±0,0007
Todas	10	0,12	0,59	4,14±0,05	0,9404±0,0007
	850	0,12	0,06	4,30±0,04	0,9439±0,0007
	2000	0,12	0,04	4,32±0,04	0,9420±0,0007

Tabla 6: resultados obtenidos para las incorporaciones de los diferentes conjuntos de nuevas variables de entrenamiento en la segunda fase, habiendo aplicado previamente la primera. El factor de trigger viene incorporado en la eficiencia.

Estableciendo el corte de BDT y BDT2 dado para este caso se obtiene el espectro mostrado en la figura 12.

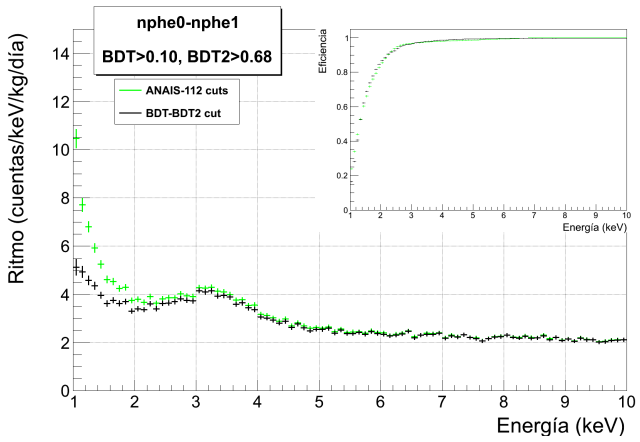

Figura 12: espectros del ritmo y la eficiencia en el intervalo de 1 a 10 keV aplicando los cortes de ANAIS (verde) y los de BDT-BDT2 (negro) incorporando nphe0-nphe1 y trabajando con 10 árboles.

Los resultados obtenidos muestran que el añadir las diferentes variables no supone una ventaja significativa frente al entrenamiento establecido previamente.

4.4. Algoritmo de una sola fase

Finalmente, se va a analizar el resultado obtenido mediante un algoritmo de entrenamiento de una sola fase. Para ello se utilizarán como variables de entrenamiento tanto las que describen la forma del pulso (presentadas en la sección 3.1) como su simetría en los PMTs (de la sección 3.2). Para las poblaciones de entrenamiento se utilizarán de nuevo la calibración como población de señal y el fondo como población de ruido. En esta última aparecen sucesos "malos" distinguibles tanto por la forma del pulso generado como por la asimetría detectada. Como ya se ha comentado, en el entrenamiento solo se utilizan los sucesos correspondientes al intervalo de 1 a 2 keV, pues en este predominan dichos sucesos. En este caso, solo se realizará un corte sobre la variable BDT asignada en una única fase, la cual se denominará, para evitar confusiones con la primera fase del algoritmo anterior, BDTsingle.

Puesto que ha vuelto a cambiar el número de variables de entrenamiento utilizadas, se van a volver a estudiar los mismos conjuntos de árboles que en la sección 4.1.1. El método de análisis será también el mismo. En la figura 13 se puede observar cómo evoluciona el valor del ritmo de fondo del detector D0 entre 1 y 2 keV para diferentes cortes en la variable BDTsingle utilizando como parámetro de control la eficiencia media entre 1 y 10 keV, con valor límite en 94,06 %. Ahora, los conjuntos 50, 100 y 500 árboles ya proporcionan la mayor reducción del fondo obtenida, por lo que se sigue trabajando con estos. Cabe mencionar la diferencia con la primera fase del algoritmo anterior, en la que se requerían hasta 850 árboles. Hay que tener en cuenta que ahora, aunque el número de variables es mayor, también se dispone de un criterio de clasificación más amplio, pues incluye tanto los aspectos relacionados con la forma del pulso como con la asimetría de este. Por ello, el algoritmo puede requerir un menor número de cortes para obtener una clasificación similar.

Además, como se ha visto en la sección 4.1.2, la segunda fase ya resultaba efectiva con 10 árboles.

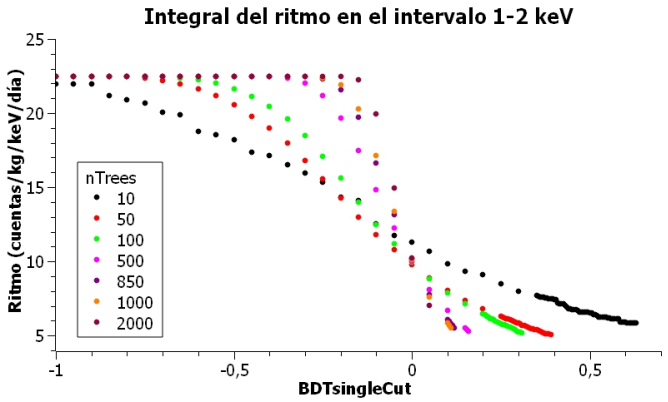

Figura 13: representación del ritmo medio entre 1 y 2 keV del detector D0 para diferentes valores de corte de BDT, trabajando con una sola fase y un límite de eficiencia media entre 1 y 10 keV de 94,06 %.

En la figura 14 se puede observar cómo se distribuyen los sucesos, tanto de calibración como los de fondo del detector D0, respecto a la variable BDTsingle definida en este caso para los tres números de árboles escogidos (rojo). También se presenta la selección de aquellos que satisfacen los cortes de ANAIS (azul). De entrada, el algoritmo parece resultar bastante efectivo, agrupando los sucesos seleccionados por ANAIS en valores de BDTsingle elevados.

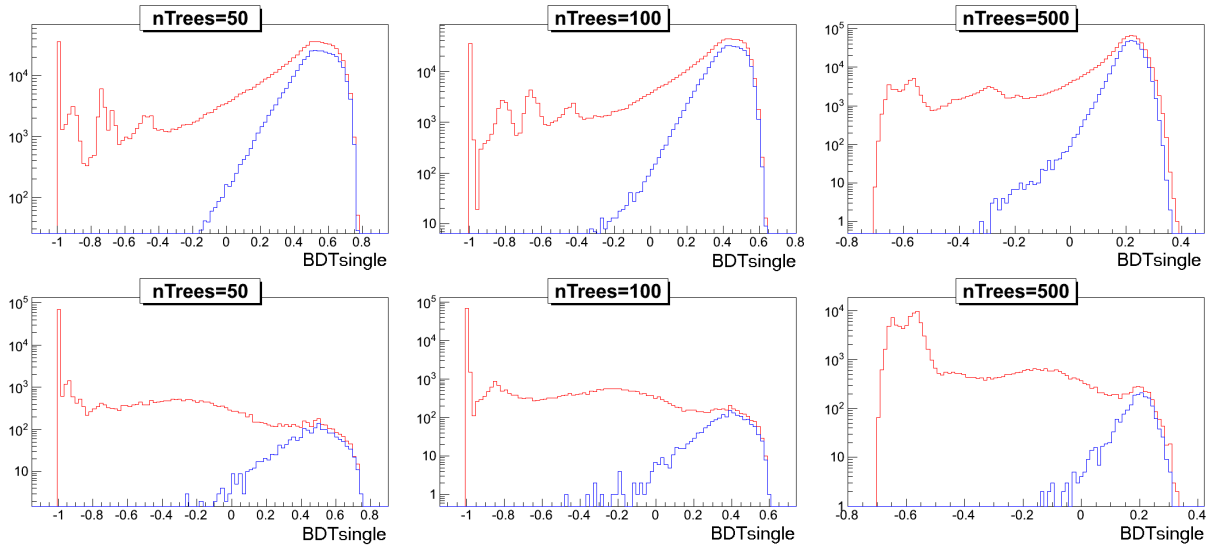

Figura 14: histogramas de la distribución de sucesos de calibración (arriba) y fondo del detector D0 (abajo) respecto a la variable BDTsingle del algoritmo constituido por una sola fase con 50, 100 y 500 árboles (de izquierda a derecha). En azul se representa la distribución de sucesos con PSVar<3 y n0,n1>4.

A continuación, se realiza un análisis del corte de BDTsingle más preciso, mediante el cuál se adquieren los valores presentados en la tabla 7. Estos resultados muestran como, trabajando sola-

mente con una fase, se puede obtener un método de discriminación igual de eficaz que el dado por dos fases. Además, este puede funcionar con tan solo 50 árboles. Se ha obtenido así un método igualmente efectivo, que supone una gran mejora en cuanto al tiempo de computación y que simplifica su aplicación al tener que aplicar un corte sobre una sola variable.

nTrees	**BDTsingleCut**	**Ritmo 1-2 keV (cuentas/kg/keV/día)**	**Eff.media**
50	0,413	3,93±0,04	0,9432±0,0007
100	0,326	4,03±0,04	0,9412±0,0007
500	0,164	4,11±0,04	0,9410±0,0007

Tabla 7: resultados obtenidos trabajando con el corte de la variable BDTsingle asignada por el algoritmo entrenado por una sola fase que minimiza el ritmo medio entre 1 y 2 keV manteniendo una eficiencia media no inferior a 94,06 % entre 1 y 10 keV .

No obstante, hay un detalle que merece la pena estudiar. En la figura 15 se muestra la distribución de los sucesos de calibración según su energía frente a la variable asignada por cada uno de los algoritmos utilizados, de izquierda a derecha: BDT de la primera fase, BDT2 de la segunda fase y BDTsingle. En esta se aprecia que, para BDT y BDT2, la distribución de sucesos queda recogida bajo un valor máximo aproximadamente constante. Sin embargo, en el caso de BDTsingle, el valor máximo hasta el que se concentran la mayor parte de sucesos varía apreciablemente, aumentando por debajo de 4 keV. Esta dependencia energética en la distribución produce que, para obtener un valor de la eficiencia igual al que se tendría con una distribución regular (mismo BDTsingle máximo), no se pueda establecer un corte tan restrictivo en el rango de 1 a 2 keV, ya que, aunque la eficiencia en este intervalo siga siendo suficientemente elevada, para energías mayores se habrán eliminado muchos sucesos de calibración y la eficiencia quedará demasiado reducida. Puesto que se está trabajando con la eficiencia media entre 1 y 10 keV, el corte de BDTsingle queda limitado, impidiendo una mayor discriminación de sucesos en el rango de 1 a 2 keV.

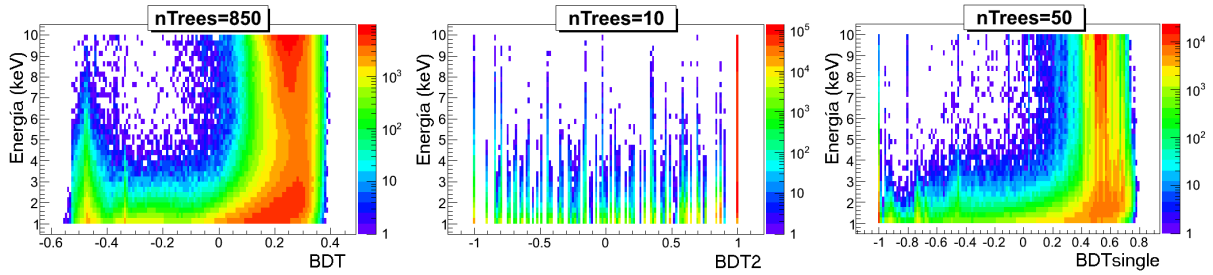

Figura 15: distribuciones de los sucesos de calibración según su energía frente a la variable BDT (primera fase, 850 árboles), BDT2 (segunda fase, 10 árboles) y BDTsingle (50 árboles). Se aplica un código de colores para indicar la concentración de sucesos.

Para evitar este inconveniente se buscan valores de corte de BDTsingle para diferentes intervalos de energía. En los apartados anteriores se ha podido observar que en el espectro de la eficiencia los cambios entre un bin y otro se vuelven apreciables en el rango de 1 a 2 keV. Esto puede producir que, al trabajar independientemente sobre este intervalo de energía, aparezca una discontinuidad

acusada en el espectro de la eficiencia general (entre 1 y 10 keV). Con el fin de evitarlo se decide que el análisis sea más preciso en el intervalo de 1 a 2 keV, trabajando con subintervalos de 0,1 ó 0,2 keV.

Para cada intervalo analizado se establece como parámetro de control la eficiencia media dada por los cortes de ANAIS en dicho intervalo (véase el Anexo III). De esta forma, se obtienen los valores de corte dados en la tabla 8.

Intervalo (keV)	**BDTsingleCut**
1,0-1,1	0,513
1,1-1,2	0,498
1,2-1,3	0,484
1,3-1,4	0,472
1,4-1,6	0,460

Intervalo (keV)	**BDTsingleCut**
1,6-1,8	0,442
1,8-2,0	0,432
2,0-3,0	0,402
3,0-4,0	0,402
4,0-10,0	0,392

Tabla 8: valores de BDTsingleCut establecidos para cada intervalo de energía del espectro del fondo promedio de los 9 detectores aplicando la eficiencia media dada por ANAIS en cada uno de ellos como parámetro de control.

4.4.1. Resultados

Aplicando los cortes presentados en la tabla 8 se llega al espectro mostrado en la figura 16, donde se puede comprobar cómo la eficiencia dada por el corte en BDTsingle se mantiene cercana a la establecida por ANAIS a lo largo de todo el rango de energía.

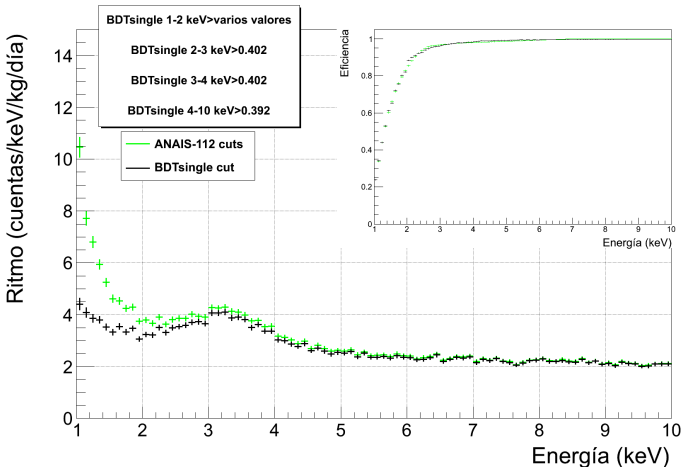

Figura 16: espectro del ritmo y de la eficiencia promedios para los 9 detectores en el intervalo de 1 a 10 keV tras aplicar los cortes de ANAIS (verde) y los de BDTsingle (negro), habiendo trabajado con 50 árboles.

En la tabla 9 se presenta el ritmo entre 1 y 2 keV alcanzado con la discriminación establecida. Siendo este el valor más reducido que se ha conseguido a lo largo del trabajo. De esta forma, el algoritmo aplicado en esta sección resulta el más eficiente, tanto por el menor tiempo de computación y análisis que requiere (pues solo se trabaja con una variable, BDTsingle), como por los resultados

que proporciona. Tras aplicar el corte en BDTsingle, el ritmo en la región entre 1 y 2 keV se reduce en un 36,8 % con respecto al obtenido por el análisis estándar de ANAIS-112. En estas condiciones, la desviación del resultado con respecto al modelo de fondo es del -26,9 %. Como ya se ha mencionado, esta diferencia puede deberse a sucesos de ruido que todavía sobreviven a los protocolos de filtrado implementados, o a otras fuentes de fondo no consideradas en el modelo.

Intervalo (keV)	**Ritmo (cuentas/kg/keV/día)**	**Eff.media**	**Diferencia del ritmo con ANAIS-112**
1-2	3,64±0,05	0,5870±0,0007	36,8 %

Tabla 9: ritmo y eficiencia medios en el intervalo de 1 a 2 keV al aplicar los diversos cortes de BDTsingle calculados y presentados en la tabla 8. Se muestra también la disminución respecto al ritmo dado por los cortes de ANAIS-112. La eficiencia incluye el factor de trigger.

5. Aplicación de los resultados

Tras el trabajo presentado se ha concluido que la mejor opción es operar con un algoritmo de una sola fase compuesta por 50 árboles, aplicando diferentes cortes según el intervalo de energía. En la figura 17 se puede observar la distribución de sucesos entre 1 y 2 keV respecto a la variable Asyarea (relacionada con la asimetría del pulso) en rojo, y la de aquellos seleccionados mediante los cortes estándar de ANAIS (azul) y BDTsingle (verde). Se puede comprobar que, tras aplicar ambos cortes, la distribución de sucesos de fondo respecto a la variable Asyarea es simétrica, como se esperaría para los sucesos de centelleo. Además, se puede observar que el corte BDTsingle conlleva una reducción mayor de los sucesos entre 1 y 2 keV.

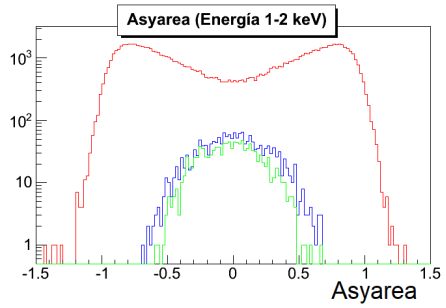

Figura 17: distribución de la variables Asyarea de la PT del detector D0 (rojo) con su selección mediante los métodos de ANAIS (azul) y BDTsingle (verde) para el rango de energía de 1 a 2 keV.

Resulta interesante analizar cómo mejora este nuevo método de selección de sucesos la sensibilidad del experimento ANAIS-112 a la señal de modulación anual de la materia oscura. La sensibilidad viene determinada por el estimador de la desviación estándar de la amplitud de modulación, $\sigma(\hat{S}_m)$. Suponiendo que el fondo permanece constante, $\sigma(\hat{S}_m)$ se puede calcular utilizando la expresión dada en el artículo [12]:

$$\sigma_{\hat{S}_m} = \sqrt{var(\hat{S}_m)} = \sqrt{\frac{2 \cdot \langle B/\epsilon \rangle}{\Delta E \cdot M \cdot T_M}} \tag{8}$$

donde ΔE es la anchura del intervalo energético considerado, M la masa total del detector (112,5 kg), T_M el tiempo de medida (considerado 5 años al ser este el tiempo esperado de trabajo de ANAIS-112) y $\langle B/\epsilon \rangle$ el valor promedio del cociente entre el ritmo de fondo y la eficiencia correspondiente por bin en el intervalo de energía considerado.

Los resultados obtenidos tras la aplicación de los cortes de ANAIS y los de BDTsingle quedan recogidos en la tabla 10, mostrando la mejoría dada por el uso de estos últimos. Como cabía esperar, es al trabajar en el intervalo 1-6 keV cuando aparece una reducción mayor en la incertidumbre, ya que la principal disminución del fondo se da entre 1 y 2 keV.

Intervalo (keV)	$\sigma_{\hat{\mathbf{S}}_\mathbf{m}}$ **ANAIS (cuentas/kg/keV/día)**	$\sigma_{\hat{\mathbf{S}}_\mathbf{m}}$ **BDTsingle (cuentas/kg/keV/día)**	**Diferencia**
1-6	0,0032	0,0028	12,50 %
2-6	0,0029	0,0028	2,99 %

Tabla 10: valores de la desviación estándar de la amplitud de modulación calculados tras aplicar los métodos de corte de ANAIS y BDTsingle.

6. Conclusiones

La selección de sucesos de centelleo es un proceso de gran importancia en el experimento ANAIS-112, pues la señal esperada de modulación anual de la materia oscura se produciría en la región de baja energía (1-6 keV), la cual está dominada por sucesos de ruido. En este trabajo se ha aplicado una herramienta de aprendizaje automático a esta tarea mediante un algoritmo BDT. Buscando su optimización se han aplicado diversas modificaciones, comprobando su eficacia y, en base a esta, descartando o manteniendo dichos cambios. Aunque se ha comenzado aplicando un algoritmo constituido por dos fases, ha resultado ser una mejor opción implementarlo en una sola, tanto por el tiempo de operación que requiere como por los resultados obtenidos, a pesar de requerir un corte variable en función de la energía de los sucesos.

El método propuesto presenta claras diferencias con respecto al utilizado por ANAIS, suponiendo una apreciable reducción del ritmo de fondo (de ≈37 %) en el intervalo de 1 a 2 keV manteniendo la misma eficiencia en ambos métodos. Esta reducción se traduce en un aumento de la sensibilidad a la señal de modulación anual de materia oscura superior al 10 %.

Tras haber comprobado el buen funcionamiento del algoritmo se presentan diferentes líneas a seguir. Por un lado, se puede continuar desarrollando el algoritmo, pues, ahora que se ha establecido el uso de una sola fase, podrían incorporarse las modificaciones aplicadas en el caso de dos fases para comprobar si mejora el resultado: separar el entrenamiento por detectores, añadir más variables de entrenamiento o incluso definir nuevas variables para los sucesos. Por otro lado, se ha visto que el uso de aprendizaje automático resulta útil en la selección de sucesos por lo que podría ser interesante aplicar otros algoritmos de este campo, como por ejemplo las redes neuronales profundas, o incluso podría investigarse el uso de redes neuronales convolucionales, que permitirían trabajar con los pulsos directamente extraídos de los fotomultiplicadores de ANAIS-112 y no con variables definidas a partir de estos. Finalmente, quedaría implementar el algoritmo estudiado, establecer los cortes adquiridos sobre los datos tomados por el experimento ANAIS-112 y analizar si se obtiene en la práctica la mejora en sensibilidad predicha según el cálculo del apartado anterior.

Referencias

[1] N.Aghanim et al. "Planck 2018 Results - VI Cosmological parameters". *A&A*, vol.641, id.A6, 2020.

[2] L.Baudis. "Dark matter detection". *J. Phys. G: Nucl. Part. Phys.*, vol.43, no.4, 2016.

[3] J.Amaré, S.Cebrián, I.Coarasa, C.Cuesta, E.Garcia, M.Martínez, M.Oliván, O.Ysrael, A.Ortiz de Solórzano, A.Salinas, M.Sarsa, P.Villar, J.Villar. "Performance of ANAIS-112 experiment after the first year of data taking". *Eur. Phys. J. C*, vol.79, no.228, 2019.

[4] K. Albertsson et al, "TMVA 4- Toolkit for Multivariate Data Analysis wiht ROOT", *CERN-OPEN-2007-007*, 2007.

[5] R.Brun and F.Rademakers, "ROOT - An Object Oriented Data Analysis Framework", "Proceedings AIHENP'96 Workshop, Lausanne, Sep. 1996", *Nucl. Inst. & Meth. in Phys. Res. A*, vol.389, pp.81-86, (1997).

[6] I.Coarasa. "ANAIS-112. Searching for the annual modulation of dark matter with a 112.5 kg Na(Tl) detector at the Canfranc Underground Laboratory", sin publicar.

[7] G.F.KNOLL. *Radiation Detection And Measurement* 3rd ed. New York: John Wiley & Sons Inc, 2009.

[8] R.Bernabei et al. "The DAMA project: Achievements, implications and perspectives", *Prog. Part. Nucl. Phys.*, vol.114, 2020.

[9] J. Amaré et al, "Annual modulation results from three-year exposure of ANAIS-112", *Phys. Rev. D*, vol.103, no.10, 2021.

[10] J.Amaré, S.Cebrián, I.Coarasa, C.Cuesta, E.Garcia, M.Martínez, M.Oliván, O.Ysrael, A.Ortiz de Solórzano, A.Salinas, M.Sarsa, P.Villar, J.Villar. (2019). "Analysis of backgrounds for the ANAIS-112 dark matter experiment". *Eur. Phys. J. C*, vol.79, no.412, 2019.

[11] Y.Coadou. "Decision trees". *EPJ Web of Conferences* 4, 02003, 2010.

[12] I.Coarasa, J.Amaré, S.Cebrián et al. "ANAIS-112 sensitivity in the search for dark matter annual modulation". *Eur. Phys. J. C*, vol.79, no.233, 2019.